

# Three speech sounds, one motor action: Evidence for speech-motor disparity from English flap production

Donald Derrick<sup>a)</sup>

*The New Zealand Institute for Language, Brain and Behavior, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand*

Ian Stavness

*Department of Computer Science, University of Saskatchewan, 176 Thorvaldson Building, 110 Science Place, Saskatoon, Saskatchewan, S7N5C9, Canada*

Bryan Gick<sup>b)</sup>

*Department of Linguistics, University of British Columbia, Totem Field Studios, 2613 West Mall, Vancouver, British Columbia, V6T1Z4, Canada*

(Received 7 June 2013; revised 21 August 2014; accepted 11 December 2014)

The assumption that units of speech production bear a one-to-one relationship to speech motor actions pervades otherwise widely varying theories of speech motor behavior. This speech production and simulation study demonstrates that commonly occurring flap sequences may violate this assumption. In the word “Saturday,” a sequence of three sounds may be produced using a single, cyclic motor action. Under this view, the initial upward tongue tip motion, starting with the first vowel and moving to contact the hard palate on the way to a retroflex position, is under active muscular control, while the downward movement of the tongue tip, including the second contact with the hard palate, results from gravity and elasticity during tongue muscle relaxation. This sequence is reproduced using a three-dimensional computer simulation of human vocal tract biomechanics and differs greatly from other observed sequences for the same word, which employ multiple targeted speech motor actions. This outcome suggests that a goal of a speaker is to produce an entire sequence in a biomechanically efficient way at the expense of maintaining parity within the individual parts of the sequence. © 2015 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.4906831>]

[C YE]

Pages: 1493–1502

## I. INTRODUCTION

Here we demonstrate that the sequence of tongue tip/blade motions in the English word “Saturday” (excluding the word-initial /s/) may be produced using a single up/down arc of tongue motion, and that other variants exist, including one with separate up/down arcs of motion for each of the two flaps (short “d”-like sounds that occur as positional variants of /t/ or /d/ in some dialects of English). We use computer simulation to demonstrate that the most commonly observed tongue motion sequence can be produced using a single cycle of muscle activation and relaxation—a single *motor action*—resulting in an arc of motion that spans two flaps and the intervening rhotic vowel. These results show that, first, a particular phonemic sequence can be produced using categorically different numbers of discretely controlled motor actions, and conversely, a single motor action may span sequences ranging from one sound (i.e., a single flap) to multiple sounds (i.e., two flaps with an intervening vowel).

This research directly addresses the pervasive assumption in speech motor behavior that units of speech production bear a one-to-one relationship to kinematically transparent speech motor actions. This assumed parity shows up in theories despite widely varying views among researchers concerning the definition of a speech motor action, as well as their matching units of speech production (e.g., see Chomsky and Halle, 1968; Meyer and Gordon, 1985; Perkell *et al.*, 2000; Browman and Goldstein, 1986, 1989, 1992).

There has long existed suggestive evidence pointing away from such parity, as in the variable contributions of jaw, lower lip, and upper lip movement in different tokens of the lip closure sequence (Folkins and Abbs, 1975), and Lisker and Abramson’s (1964) observation that American English and Persian speakers pre-voice some initial unaspirated stops, but not others, for the same word and context. Such examples, however, may simply be part of a gradient spectrum of production variation. In order to effectively demonstrate that speech production violates parity, categorical examples are needed.

Perhaps the best-known case of apparently categorical variation is reported by Delattre and Freeman (1968), who describe eight categorical variants of the English rhotic (hereafter “R” in this paper). For the purposes of this paper, we focus on two broad types of variants, those with the tongue tip-up [ɹ], which include tip-up bunched and retroflex

<sup>a)</sup>Author to whom correspondence should be addressed. Also at The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, New South Wales 2751, Australia. Electronic mail: donald.derrick@gmail.com

<sup>b)</sup>Also at Haskins Laboratories, New Haven, Connecticut 06511.

R, and those with the tongue tip-down [ɹ̩]. These variants extend across speakers based on dialect, and within speaker based on phonological context (see Westbury *et al.*, 1999; Stavness *et al.*, 2012). While English rhotics provide a remarkable case of conditioned categorical variation, this kind of variation has not generally been observed in the same word and phonological context, and it has not provided a challenge to assumptions of parity between motor actions and speech sounds.

Describing a case of within-context variation, Derrick and Gick (2011) identify four qualitatively different ways of producing flap/tap variants (hereafter “T” in this paper) that differ based on how the tongue tip approaches, contacts and leaves the alveolar ridge of the palate: An upward “up-flap” motion ([ɹ̩]), a downward “down-flap” motion ([ɹ̸̩]), an up-down “alveolar tap” motion ([ɹ̩̩̩]), and a front-back “postalveolar tap” motion ([ɹ̩̩̩̩]). They find that a single speaker will use different T variants for the same speech sound produced in the same word and sentence context, showing that one speech sound may correspond to multiple apparent speech motor actions.

The present paper focuses on evaluating the plausibility of the converse point: That a single speech motor action may encompass multiple speech sounds, providing further evidence against an assumption of parity. We test this by identifying a sequence in which multiple T motions might represent a single larger arc of motion, and then using simulations to see whether that pattern of motion might result from one underlying set of muscle activations.

We chose sequences such as that in the word “Saturday” because such sequences enable us to observe the interplay between these two cases of extreme categorical variation: T and R. Casual observation of existing x ray films reported in Cooper and Abramson (1960) reveals that three of the four talkers in that dataset produce the two consecutive T’s in the word “Saturday” as an up-down flap sequence, as shown in Fig. 1. However, based on the frequency of T variants observed by Derrick and Gick (2011), considering all kinematically plausible combinations, we should expect (all else being equal) an up-down flap sequence to occur only 28.5% of the time. We first corroborate this observed overrepresentation of the up-down flap sequence with a more substantial and controlled study, and second, consider what additional factors may play into a preference for this up-down sequence.

We hypothesize that in North American and other varieties of English, the word “Saturday” typically involves two opposite movements of the tip of the tongue for the sequence of T consonants: an upward-rearward motion followed by a

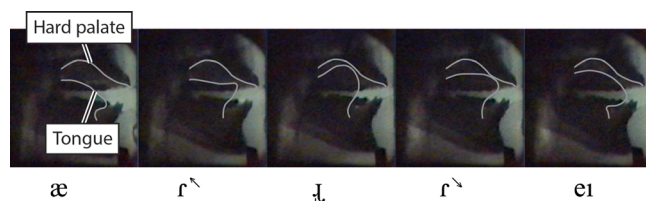


FIG. 1. (Color online) X-ray data showing a production of Saturday as [sæɹ̩̩̩̩r̩̩̩̩eɪ]. (Data from Cooper and Abramson, 1960.)

downward-forward motion, giving an up-down [ɹ̩̩̩̩ɹ̩̩̩̩] sequence. We argue that what is attractive about this up-down sequence from a speech production point of view, and would account for this overrepresentation, is that there is one speech motor action that encompasses the entire up-down tongue movement sequence spanning three segments ([ɹ̩̩̩̩ɹ̩̩̩̩]) in “Saturday.” That is, the entire sequence may be realized as a single, cyclic motor action where the upward movement is produced through muscle activation and the downward movement occurs passively due in large part to two factors: Gravity and elasticity.

Considering gravity, the human neuromuscular system partly compensates for the effects of gravitational load on speech; thus, jaw motion during speech differs somewhat based on whether a speaker is prone (face down) or supine (face up) (Shiller *et al.*, 1999). The results from the research of Shiller *et al.* (1999) also show that tongue motion does not entirely compensate in place of jaw motion, as evidenced by differences in measurements of F1 and F2 during vowel production in prone and supine position. The evidence therefore demonstrates that speech motor actions depend on assumptions about the direction of gravity.

Considering elasticity, Perrier *et al.* (2003) have provided experimental and two-dimensional finite element method (FEM) vocal tract simulation-based evidence that tissue elasticity factors in the motions of vocal tract articulators during the production of velar stops. FEM is a well-known computational technique for calculating the effect, or distribution, of stress and strain within a structure to which forces are applied, and is therefore useful for modeling the biomechanics of muscle, cartilage and bone. In their example, much of the forward looping pattern of velar stop production in vowel-consonant-vowel (VCV) sequences is based on the anatomical structure of the tongue such that planning may be based on target sequence as much as or more than trajectory motion. This suggests that the planning system incorporates information about the structure and elasticity of the anatomy.

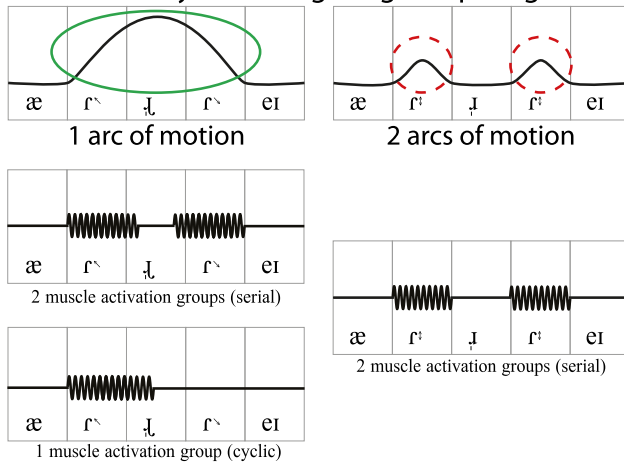
On the basis of the potential effects of gravity and elasticity on articulator motion and planning, it is reasonable to expect that both forces contribute to the production of [ɹ̩̩̩̩] by passive lowering of the tongue tip from an initial high position above the alveolar ridge. This hypothesis, if correct, allows for one speech motor action to encompass the production of three speech sounds and two directions of motion spanning a syllable boundary.

Figure 2 shows how an up-down sequence of tongue tip movements in the word Saturday (left, [ɹ̩̩̩̩ɹ̩̩̩̩]) might be produced using either a group of active muscle activations each for [ɹ̩̩̩̩] and [ɹ̩̩̩̩], or one group of active muscle activations for the [ɹ̩̩̩̩], followed by muscle relaxation for the [ɹ̩̩̩̩]. In contrast, an alternative production strategy using a sequence of two taps (right, [ɹ̩̩̩̩ɹ̩̩̩̩]) will always require two distinct sets of muscle contractions.

## A. Hypothesis

We hypothesize that a single motor action may govern multiple observable kinematic events spanning multiple speech segments. This hypothesis leads to two predictions:

### 'Saturday' - tracking tongue-tip height



### Potential underlying muscle control

FIG. 2. (Color online) Schematic of possible underlying patterns of muscle contractions for production of the tongue tip motions in the word “Saturday.”

- (1) The word “Saturday” will usually be produced with a  $[r̥ɹ̥]$  sequence. That is, there will be more instances of tip-up  $[ɹ̥]$ , as opposed to tip-down  $[ɹ]$  for the rhotic vowel in the word “Saturday” as compared with a similar word containing a rhotic vowel without flanking T’s, such as “peppermint” (which will have more instances of tip-down  $[ɹ]$ ).
- (2) Gravity and myoelasticity can be demonstrated, within a biomechanically realistic vocal tract simulation, to passively complete a  $[r̥]$  closure and complete it fast enough to produce a T instead of a stop (i.e., in about 10 ms).

Below we present our experiments, followed by our computer simulations, in the same order as the introduction above.

## II. EXPERIMENT

The use of ultrasound imaging to look at midsagittal slices of the tongue (B-mode), along with three one-dimensional slices that cut through the tip and blade of the tongue (M-mode), can provide information about tongue-tip motion in rapid sequences. Using a narrow transducer placed against the skin near the angle of the neck, B-mode ultrasound provides a low speed image [30 frames per second (fps)] of the overall shape of the midsagittal surface of the tongue from the root to the tip. M-mode ultrasound provides high-speed trajectories, dependent upon equipment and settings, of the direction of tongue motion through fixed cross-sections in the vocal tract.

We expect the R in the word “Saturday” to be realized as  $[ɹ̥]$  more often than the R in the word “peppermint.” We also expect most instances of the first T variant  $[r̥]$  to be followed by  $[ɹ̥]$ , whereas we would expect most instances of the first T variant  $[r̥^1]$  to be followed by  $[ɹ̥]$ . Similarly, we expect most instances of  $[ɹ̥]$  to be followed by  $[r̥^1]$ , and most instances of  $[ɹ]$  to be followed by  $[r̥^1]$ . As a result of

the expected strong preponderance of  $[ɹ̥]$  as described above, we expect that most of the sequences in “Saturday” will be  $[r̥ɹ̥]$  sequences, as per our single motor action hypothesis. Most of the rest should be  $[r̥^1ɹ̥]$  sequences, and so produced by two distinct speech motor actions.

## A. Experiment methods

The experimental methods below are described in [Derrick and Gick \(2011\)](#). Twenty-six native speakers of North American English between the ages of 18 and 40 participated in the study. Eight of the participants (participants 1, 7, 11, 19, 20, 22, 24, and 25) consistently produced complete stop closures instead of T variants during read speech, leaving 18 participants (ten males and eight females). All participants had normal speech and reported normal hearing. Participants were seated in a customized American Optical Co. model 507-a (1953) ophthalmic chair with a two-cup rear headrest adjusted to contact the base of the skull just above the neck.

A UST-9118 EV 180 electronic curved array ultrasound probe was placed under the chin. The probe has a variable frequency range of 3–9.0 MHz with an average  $\mu$  slice thickness of the tissue viewed with this probe of approximately 3 mm ([Medicines and Healthcare Products Regulatory Agency, 2004](#)). The probe was attached to an Aloka ProSound SSD-5000 ultrasound machine connected via s-video cable (marked video IN) to a Canopus ADVC-110 advanced digital video recorder.

A Sennheiser MKH-416 short shotgun microphone was mounted on a microphone stand and aimed at the participant about 30 cm away from the mouth. The microphone was plugged into a M-Audio DMP3 “Audio-buddy” pre-amplifier via XLR balanced cable and out with an unbalanced RCA cable to the Canopus card to guarantee time synchronization between the ultrasound and audio output.

The ultrasound machine was set up in simultaneous B/M mode and aligned to the acoustic signal. B-mode ultrasound was used to capture two-dimensional images of the midsagittal plane of the tongue at 30 fps. The M-mode (motion mode) ultrasound provided a progressive scan of three selected one-dimensional lines accessible from an ultrasound probe. These one-dimensional M-mode lines follow the line of the palate, in the region of intercept with the blade/tip of the tongue. Because M-mode ultrasound is a progressive scan, it presents the motion data at the full capture rate of the ultrasound probe, which ranged from 60 to 100 Hz depending on the depth of the scan. While this motion is not connected to any specific flesh-point, it allows capture of the general direction of motion of the front of the tongue, which is ideal for identifying the T variants described above. At the same time, the B-mode ultrasound allows examination of the midsagittal plane of the tongue surface at 30 fps, which along with the M-mode data allowed identification of the R variants described above.

An LCD monitor was mounted on the ophthalmic chair’s monitor mount and placed in front of the participant. A computer containing the experiment stimuli presentation software was connected to the LCD monitor so that the



participant could easily read the stimuli from the screen. Stimulus tokens were selected to contain single T or sequences of two T's within consecutive syllables. Data were collected on 17 control sentences, nine sentences with 1 T, ten sentences with double T sequences, and two sentences with triple T sequences, for a total of 38 unique sequences. The sentences were randomized for each of 12 blocks, giving a total of 456 stimulus sentences. The stimuli were presented using the psychological experiment presentation tool PXlabRT (Irtel, 2007) set to present stimuli such that each sentence was displayed on an LCD screen for 2.2 s for a total of 12 blocks. The software automatically paused the experiment after the first six blocks (9 min) to allow participants to swallow some water or take a short break if needed. The 12 blocks were presented in set order, but the entire set of 38 sentences was randomized for each block. The present report is based on data collected as part of the larger experiment described here, but with particular focus on the two tokens, "We have Saturday off" and "We have peppermint now."

Participants were asked to repeat "ta" at least ten times rapidly in order to record tongue motion speed and to provide data for audio synchronization. Participants were then asked to repeat sentences containing T sequences while the ultrasound machine was configured to match the size and shape of their head and tongue. The experiment software was then activated and experiment data were recorded as described above. Participants were then asked to produce the 38 stimuli, in randomized 12 blocks (with a short break between block six and seven), for a total of 456 stimuli. Each block took 9 min, for a total of 18 min recording time.

Data were recorded directly onto a Macbook via the Canopus card, and the audio was extracted from the DV recordings. Audio-video synchronization was confirmed

using the sequences of acoustic transients from the alveolar stop releases in the spoken sequences of "ta" with tongue dropping gestures associated with the same. The Canopus card's audio and video synchronization were consistently within one frame throughout the experiment, requiring no special post-production synchronization.

The acoustic signal was labeled and transcribed in Praat (Boersma, 2001), with attention to identifying segment boundaries and the acoustic low amplitude point (center) of each T. Data were then imported into ELAN, a tool for annotating audio and video recordings simultaneously (Sloetjes and Wittenburg, 2008). The tongue positions of each R were identified by examining the tongue position at vowel mid-points, as seen in the B-mode ultrasound data, and coded as to whether the rhotic vowel was [ɹ] or [ɹ̥].

The T closure times were identified as the point of lowest acoustic amplitudes (Zue and LaFerriere, 1979). The T variants themselves were identified using both the B-mode data, and more importantly the M-mode ultrasound to track the motion of the tongue tip and blade. As noted above, M-mode provides a one-dimensional progressive scan of motion along chosen intersect lines. When M-mode intercept lines are aligned as seen in the top portion of Fig. 3, the surface of the tongue tip/blade as it crosses the intersect shows up as a white line; the white line is higher when the tongue tip/blade is high and back, and lower when the tongue tip/blade is low and front. The T variants are identified by first examining the B-mode video just before, during, and after T contact to see overall tongue motion. The identification is confirmed by examining M-mode data from a couple of frames ahead of the flap contact, focusing on the M-mode data adjacent to the leading edge, as identified by the thick black lines, and highlighted as the *area of interest* in Fig. 3.

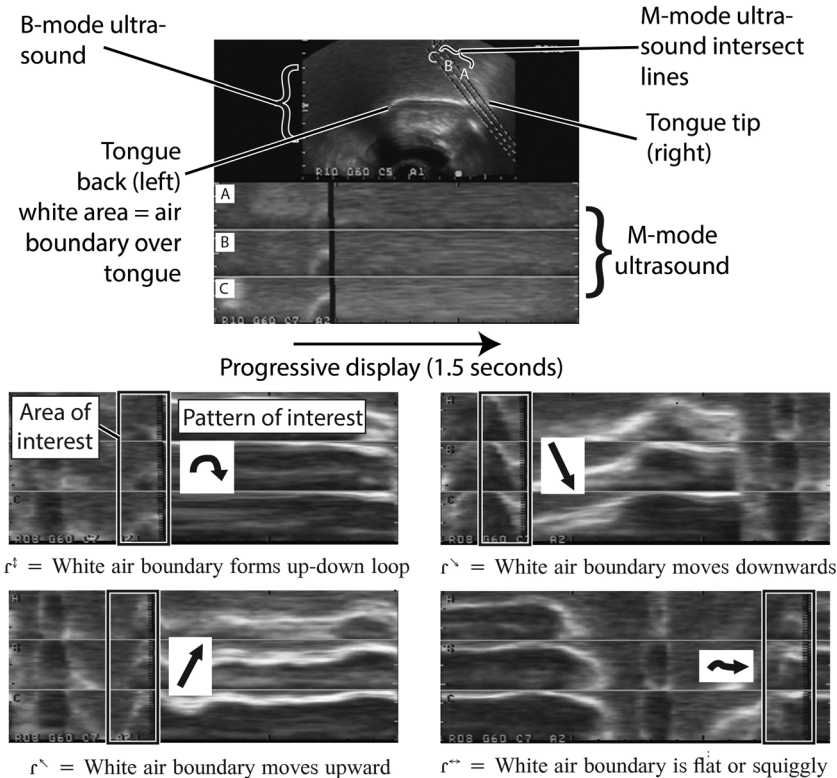
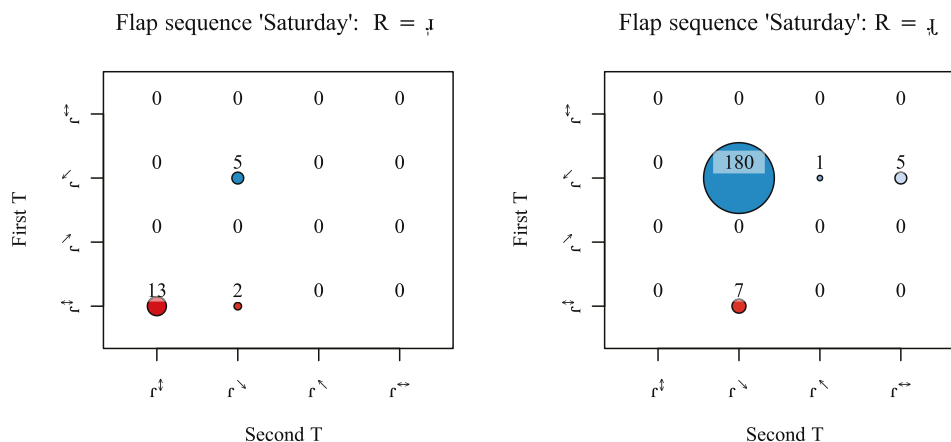


FIG. 3. Schematic of B/M mode ultrasound with visualization of the technique for identifying T variants through M-Mode.



Within the M-mode data, there are four *patterns of interest* illustrated in Fig. 3: Alveolar taps ( $[r^{\uparrow}]$ ) are identified by an up-down loop centered around the acoustically identified time of contact. Down-flaps ( $[r^{\searrow}]$ ) are identified by a downward motion of the white air boundary. Up-flaps ( $[r^{\nearrow}]$ ) are identified by an upward motion of the white air boundary. Last, postalveolar taps ( $[r^{\leftrightarrow}]$ ) are identified by a flat or slightly squiggly horizontal white air boundary, higher than the typical up-down loop of an  $[r^{\uparrow}]$ .

Statistical analysis was completed in R (R Core Team, 2013) using Wilcoxon signed-rank tests, which provide a conservative replacement for paired Student *t* tests in data where normality cannot be assumed.

## B. Experiment results

Comparing the frequency of R variants in the words “Saturday” vs “peppermint” reveals that speakers are more than twice as likely to produce [ɹ] for “Saturday” (193 of 213 tokens) than for “peppermint” (71 of 210 tokens). Wilcoxon signed-rank tests were performed, and for each of the two R variants, the percentage of productions matching that tongue tip position based on whether the word in question is “Saturday” or “peppermint” is compared. The results confirm a significant difference between the two words ( $V = 147.5, p < 0.001$ ).

### C. First T in “Saturday”

Most of the first T variants in “Saturday” were  $[r^{\searrow}]$ , or 191 out of 213. Of these, 186 were followed by  $[j]$ . In contrast, of the 20 tokens of “Saturday” with  $[j]$ , 15 of the first T variants were  $[r^{\uparrow}]$ . Wilcoxon signed-rank tests were performed on the data summarized in Fig. 4 using T variant as the independent variable, and R variant as the dependent variable. For each of the four T variants, the percentage of productions matching that T variant based on the R variant in “Saturday” were compared. As expected from the descriptive statistics in Fig. 4, the results are significant for  $[r^{\searrow}]$  such that the following R variant is significantly more likely to be an  $[j]$  ( $V = 1$ ,  $p = 0.001$ ). There were not enough instances of  $[r^{\uparrow}]$  for the test to demonstrate that they were more likely to occur with  $[j]$ .

### D. Second T in “Saturday”

Of the 193 tokens of “Saturday” produced with [ɹ], fully 187 ended with [r↘], and only six ended with another T variant. In comparison, of the 20 tokens of “Saturday” with a [ɹ], 13 ended with [r↑], and only seven ended with [r↘]. Wilcoxon signed-rank tests were performed on the data summarized in Fig. 4 using T variant as the independent variable, and R variant as the dependent variable. For each of the four T variants, the percentage of productions matching that T variant based on the R variant in “Saturday” were compared. The results are significant for [r↘] ( $V=0$ ,  $p < 0.001$ ). Again, there were not enough instances of [r↑] to demonstrate that they were more likely to occur with [ɹ].

### E. TRT sequences

The results also show that, of the 213 T sequences among the 18 participants of this study, 180 of them were  $[r \downarrow \downarrow r]$  sequences, representing 84.5% of the sequences, as seen in Fig. 4.

### III. SIMULATION

Disentangling and ranking the influence of the various factors that contribute to tongue movement in speech (muscle forces, tissue elasticity, gravity, etc.) is difficult to do with experimental measurement alone. Computer simulations of biomechanical systems are well suited for detailed analysis of the factors that underlie movement because such simulations describe how the forces within the system interact to generate movement. For our simulation analysis, we used a biomechanical simulation toolkit, ArtiSynth (UBC, Canada, version 2.9, [www.artisynth.org](http://www.artisynth.org)), that has been specifically designed for modeling the human vocal tract (Fels *et al.*, 2003; Lloyd *et al.*, 2012). We used a three-dimensional model of the jaw-tongue-hyoid-palate (JTHP) that includes muscles forces, elasticity, and gravity and accounts for dynamic coupling between the articulators (Stavness *et al.*, 2012). The jaw-tongue-hyoid-palate (JTHP) model was built from reference tongue (Buchallard *et al.*, 2009) and jaw (Hannam *et al.*, 2008) models that were adapted to fit a computed tomography scan of a single subject (Stavness *et al.*, 2012). The model is pictured in Fig. 5.

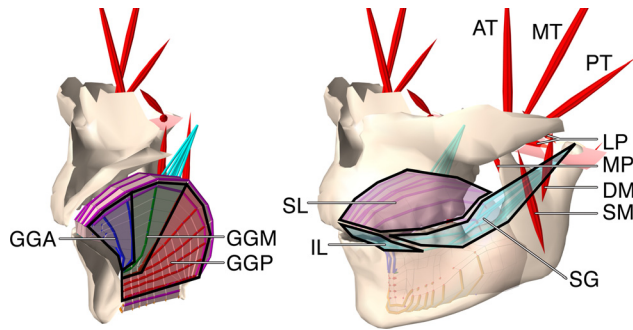


FIG. 5. (Color online) (Left) Cutaway and (right) oblique views of the JTHP model. Jaw muscles are shown as lines and include the anterior/middle/posterior temporalis (A/M/PT), lateral pterygoids (LP), medial pterygoid (MP), deep/superficial masseter (D/SM), and posterior/anterior belly of the digastric (not shown). Tongue muscles are denoted by shaded areas and include the anterior/middle/posterior genioglossus (GGA/M/P), superior/inferior longitudinal (S/IL), styloglossus (SG), as well as the mylohyoid, geniohyoid, hyoglossus, transversus, and verticalis muscles (not shown).

Tongue elasticity is represented in the model by the finite-element (FE) method with a non-linear, nearly incompressible hyperelastic material. The elasticity properties for the material were taken from literature data in combination with mechanical testing with fresh cadaveric tongue tissue (Gérard *et al.*, 2006). These measurements were used to fit parameters in an isotropic, non-linear, hyperelastic material—a fifth-order Mooney-Rivlin material (Mooney, 1940; Rivlin, 1948),

$$W = C_{10}(I_1 - 3) + C_{20}(I_1 - 3)^2 + \frac{\kappa}{2}(\ln J)^2,$$

where the  $(\kappa/2)(\ln J)^2$  term enforces tissue incompressibility. Other terms in the Mooney-Rivlin material were omitted, i.e.,  $c_{01} = c_{11} = c_{02} = 0$ . Material coefficients were found of  $c_{10} = 1037$  Pa,  $c_{20} = 486$  Pa (Buchillard *et al.*, 2009). The model used Rayleigh damping, which is a viscous damping proportional to both tissue stiffness ( $\beta$  coefficient) and tissue mass ( $\alpha$  coefficient). Rayleigh damping coefficients were set to achieve critically damped response for the model ( $\beta = 0.03$  s and  $\alpha = 40$  s<sup>-1</sup>).

The tongue model's FE mesh includes 740 hexahedral elements with a density of 1040 kg/m<sup>3</sup> for a total tongue mass of 106 g. In the JTHP model, tongue elasticity is dynamically coupled to the jaw and contact is handled between the tongue-jaw and tongue-palate. Gravity is included in the model as a constant downward force of mass  $\times 9.81$  m/s<sup>2</sup> applied to the jaw, hyoid bone, and all of the FE nodes in the tongue model.

Muscle forces are represented in the model by a set of Hill-type muscle models (Zajac, 1988). The jaw model includes 20 Hill-type line muscles to represent the main compartments of the mandibular muscles. The tongue model includes numerous Hill-type muscle fibers embedded within the FE mesh. The FE mesh was constructed to approximate the shape of the lingual muscles (based on Takemoto, 2001); therefore, muscle fibers are embedded along the edges of the FE mesh. Muscle control is based in part on electromyography (EMG) studies of the tongue which argue for partially independent control of parts of the genioglossus (Miyawaki

*et al.*, 1975). Otherwise, due to the high dimensionality and difficulty of identifying smaller groupings of motor units that control parts of tongue muscles (see Slaughter *et al.*, 2005), the model uses the anatomical structures of muscles as control groups (11 bilateral muscle groups in total).

The simulations reported in this study are forward dynamics simulations. The inputs to the dynamic simulation are time-varying muscle activations (within the range of 0.0–1.0). At each timestep of the simulation, muscle forces are calculated based on the current muscle activations via the Hill-type muscle models, those forces are applied to the model, the model's acceleration is calculated by Newton's second law, and then the model's velocity and position are calculated by numerical integration (see Lloyd *et al.*, 2012, for a mathematical description of the simulation process). Therefore, the output of the simulation is both the time-varying muscle forces (which depend on muscle activations) as well as the kinematics of the jaw, hyoid bone, and the FE nodes of the tongue.

Previous simulations reported for the coupled JTHP model have shown plausible speech (e.g., Stavness *et al.*, 2012) and chewing (e.g., Lloyd *et al.*, 2012) motions; therefore, we believe it is suitable for our simulation needs. Here, we use this model to investigate the effect of muscle forces, elasticity, and gravity on [r̥] closure.

We expect certain muscles to participate in the formation of an [r̥] and following [ɹ], such as muscles for raising the jaw, the superior longitudinal (SL) muscle for curling up the tongue tip, the posterior genioglossus (GGP) and medial genioglossus (GGM), for advancing the tongue tip and body, and the transversus (TRANS) for narrowing the tongue and elevating the surface. We also found the styloglossus (SG) was necessary to retract the tongue sufficiently to allow the production of a (retroflex) [ɹ]. For the production of [r̥] followed by [ɹ], we expect that contracting the muscles above will lead to tongue-tip motion upward, contacting the alveolar ridge and pulling away into a tip-up (retroflex) position. Potential agonists for the [r̥] include the anterior genioglossus (GGA) and inferior longitudinal (IL) muscle for lowering the tongue tip. However, we do not expect these muscles to be needed to produce a [r̥].

We used the JTHP model to test whether the [r̥ɹ̥] sequence can be produced with muscle activations for the [r̥] only, and we used the JTHP model to create an [r̥ɹ̥] sequence via direct activation of muscles for both T's. For the active (serial, two-action) simulation, we demonstrate that the [r̥] motion into the [ɹ] can be generated with one set of muscle contractions, and that the [r̥] can be generated with the help of a second set of muscle contractions. For the passive (cyclic, one-action) simulation, we expect the [r̥] contact will occur during or just after completion of muscle activations for the [r̥], and the [r̥] contact will occur shortly after muscle deactivation. The duration between the [r̥] and [r̥] may be determined by either the strength of the initial muscle activations, or the length of time during which the muscle activations were sustained—any combination of the two should function to similar effect. The [r̥] contact in the passive simulation will



occur slower, last longer, and possibly occur at a slightly different tongue-contact point than in the active model, but still fast enough to be a flap and not a stop. For these reasons, this slower  $[r^{\searrow}]$  will be distinguishable from the active model  $[r^{\searrow}]$ . Nevertheless, we expect the differences to be subtle enough to render it difficult if not impossible to identify the differences in human experiments without EMG recordings.

## A. Simulation methods

To test the two simulation hypotheses above, we created two simulation models. Input probes were created for the JTHP model in order to simulate a  $[r^{\searrow}]$  followed by the tongue-tip position for a retroflex  $[ɻ]$ .

For both models, the jaw positioning and  $[r^{\searrow}]$  muscle activations were the same. Bilateral jaw elevators (masseter, temporalis, and medial pterygoids) were programmed to move the jaw into position for speech from 10 to 290 ms. For  $[r^{\searrow}]$  muscle activations, the SL probe was set to 0.57 standard units of activation, TRANS to 0.33, the GGP to 0.2, and GGM to 1.6. All four probes were set to activate at 50 ms, completing activation at 75 ms (for a 25 ms attack), be sustained for 100 ms, and then relax starting at 175 ms, reaching 0 at 200 ms (for a 25 ms decay). These four muscles were used to create the  $[r^{\searrow}]$  motion of the tongue tip along with the characteristic tongue shape for  $[ɻ]$ . The SG probe was also activated to 0.43 standard units, starting at 50 ms, but at 100 ms (for a 50 ms attack), activation was sustained over a 50 ms, relaxation began at 150 ms, and ended at 200 ms (for a 50 ms decay). The SG muscle was used to pull the tongue away from the alveolar ridge into a retroflex  $[ɻ]$  position. These very specific activations were generated from well-known ideas about how the tongue tip is raised, and careful heuristic tuning of the JTHP system.

The JTHP models were then run with the above input probes, and the position of the tongue tip was recorded from the beginning of activation until complete relaxation of the SL, TRANS, GGP, and GGM ( $[r^{\searrow}]$ ) probes in order to see if the tongue moved through a  $[r^{\searrow}]$  while the muscles were relaxing.

The passive simulation involved no other muscle activations. For the active model, the  $[r^{\searrow}]$  muscle activations involved two muscles. The GGA and the IL was set to activate to 0.2 standard units beginning at 175 ms, reaching full activation at 200 ms (for a 25 ms attack), and then deactivate completely by 225 ms (for a 25 ms decay). The active simulation had the  $[r^{\searrow}]$  probes activate while the  $[r^{\searrow}]$  probes were deactivating such that they reached full activation just as all the  $[r^{\searrow}]$  probes were fully deactivated.

## B. Simulation results

The results of the simulations for the active and passive models are presented below. These include the timings of  $[r^{\searrow}]$  contact, mid-point of the  $[ɻ]$ , the  $[r^{\searrow}]$  contact and mid-point of the final vowel, all in relation to the muscle activations.

## 1. Active simulation

The active (or “serial”) simulation shows that the  $[r^{\searrow}]$  achieves alveolar ridge contact for 14 ms, from the 83 ms marker to the 96 ms marker. This constitutes a suitable duration of tongue tip contact for a flap, as opposed to a stop. The tongue tip reaches its furthest distance from the alveolar ridge at 135 ms. The second contact for  $[r^{\searrow}]$  takes place between 173 and 179 ms, lasting 7 ms. The contact location of  $[r^{\searrow}]$  is posterior to and higher along the alveolar ridge than that of the preceding  $[r^{\searrow}]$ .

Flap contacts can be seen in Fig. 6. Flap contact is indicated through the ArtiSynth collision detection system, and appears as dark lines radiating from the points of collision above the light ball indicating a finite element node.

## 2. Passive simulation

The passive (or “cyclic”) simulation shows that, while the  $[r^{\searrow}]$  and  $[ɻ]$  are generated through active muscle control, the subsequent  $[r^{\searrow}]$  occurs passively during relaxation of the same muscles (i.e., as a result of the passive elasticity and gravitational forces in the model). The  $[r^{\searrow}]$  achieves alveolar ridge contact for 14 ms, from the 83 ms marker to the 96 ms marker, just as in the active model. The tongue reaches its furthest distance from the alveolar ridge at 135 ms, just as in the active model. The  $[r^{\searrow}]$  takes place between 173 and 182 ms, lasting 9 ms, constituting a slightly longer and firmer flap than in the active model. The  $[r^{\searrow}]$  contact takes place posterior to the  $[r^{\searrow}]$ , higher along the alveolar ridge, just as in the active model. The results are seen in Fig. 7.

The simulation can be programmed to provide shorter and longer retroflex  $[ɻ]$  durations based on the strength and/or length of muscle contractions. Stronger contractions lead to more pronounced retroflexion and shorter tongue-tip contact durations during the  $[r^{\searrow}]$  but are otherwise similar to the simulations above. Note that simulations with gravity turned off failed to reach targets.

## IV. DISCUSSION

The results of the experiment and simulation support the hypothesis that a single motor action may govern multiple observable kinematic events spanning multiple speech segments. The results show that, as predicted, speakers produce the  $[r^{\searrow}ɻr^{\searrow}]$  sequence for “Saturday” most of the time. There are 185  $[r^{\searrow}]$ ,  $[r^{\searrow}]$  sequences recorded out of 213 tokens for the word “Saturday,” the R in the word “Saturday” is significantly more likely to be  $[ɻ]$  (193 out of 213, or 90.6%) than the ones in the control phrase “peppermint” (71 out of 210, or 33.8%), such that fully 180 of 213, or 84.5% sequences in “Saturday” were produced as  $[r^{\searrow}ɻr^{\searrow}]$  sequences. These sequences involve a single up/down arc of motion, whereas, in contrast, the 13  $[r^{\uparrow}ɻr^{\uparrow}]$  sequences involve two separate up/down arcs of motion, one for each  $[r^{\uparrow}]$ . As expected, the up-down flap sequence is thus dramatically overrepresented in our production results.

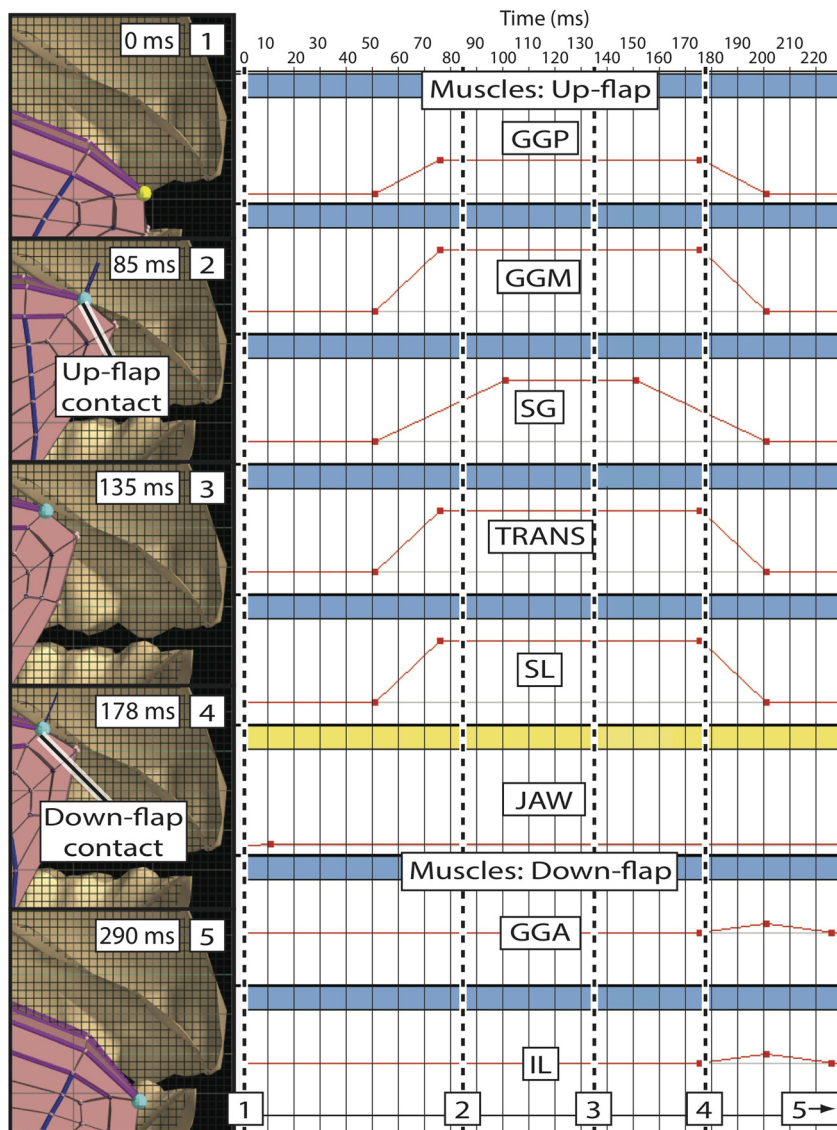


FIG. 6. (Color online) Active model: Tongue tip positions in relation to ArtiSynth muscle activations with active  $[r^{\downarrow}]$  and  $[r^{\uparrow}]$  muscle activations.

The results of the simulations further show that the effects of gravity and myoelasticity are sufficient to govern  $[r^{\downarrow}]$  production; gravitational and myoelastic forces allow the  $[r^{\downarrow}]$  contact to occur quickly enough such that total occlusion of the alveolar ridge lasts considerably less than 15 ms, which is about the maximum duration of a T contact, as opposed to that for a full stop consonant.

The active model produced the  $[r^{\downarrow}]$  in a similar fashion to the passive model, but more quickly, with a slightly shorter contact duration, and with a different tongue position after the T (more like that of a low front vowel). As a result, it is at least possible that speakers could opt to use an active  $[r^{\downarrow}]$  if the following vowel is low, and more likely to use passive  $[r^{\downarrow}]$  if the following vowel is mid or higher, as in the [ei] in “Saturday.” The shorter contact duration was not predicted in the hypothesis and may have resulted from the change in contact location generated from the active muscle contractions. Regardless, the success of the passive model demonstrates that  $[r^{\downarrow}]$  motions can potentially be produced passively, and therefore that the entire sequence  $[r^{\downarrow}\downarrow r^{\downarrow}]$  can be produced as a single, cyclic speech motor action. In comparison, the second most

commonly attested sequence of  $[r^{\downarrow}\downarrow r^{\downarrow}]$  must involve at least two speech motor actions.

That is, the arcs of motion for each  $[r^{\downarrow}]$  in a  $[r^{\downarrow}\downarrow r^{\downarrow}]$  span fewer segments than the larger arc of motion for  $[r^{\downarrow}\downarrow r^{\downarrow}]$ . While the short single  $[r^{\downarrow}]$  and longer  $[r^{\downarrow}\downarrow r^{\downarrow}]$  sequences may appear somewhat similar in overall pattern, they are quite different in duration and in the tasks they govern. The shorter  $[r^{\downarrow}]$  action may be construed as being directed at a single spatial constriction task (Saltzman and Kelso, 1987; Saltzman and Byrd, 2000) or target (Browman and Goldstein, 1986, 1989, 1992), while the larger  $[r^{\downarrow}\downarrow r^{\downarrow}]$  action involves the tongue tip and blade cycling through a motion that reaches several articulatory targets. Since these sequences can be produced by the same speaker either as a single cyclic event, or as a sequence of discrete events, it appears that the speaker’s goal is to produce the sequence in the most biomechanically efficient way as an entire sequence, as opposed to maintaining parity within the individual parts of the sequence. The loss of parity in the model, however, corresponds with a gain in the ability of the system to select from among alternative cyclic and targeted actions based on the dynamic needs of the current speech event (see Grillner, 2006; Dominici *et al.*, 2011;



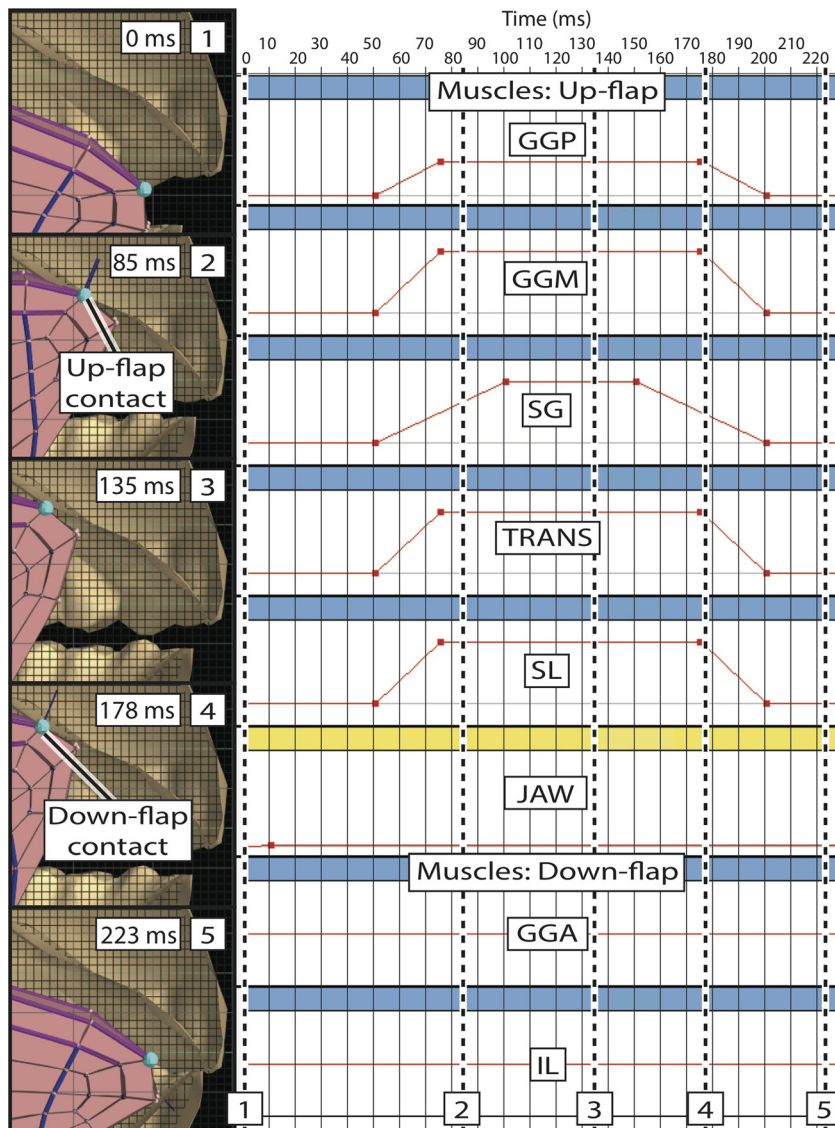


FIG. 7. (Color online) Passive model: Tongue tip positions in relation to ArtiSynth muscle activations with  $[r^{\sim}]$  muscle activations, and no  $[r^{\sim}]$  muscle activations.

Gick and Stavness, 2013). The overrepresentation of this cyclic action may be interpreted as implicating a central pattern generator or other oscillatory primitive mechanism for speech movements of this kind (e.g., see Barlow and Estep, 2006; Lund and Kolta, 2006). We expect that there exist many more such cases, including potential many-to-many disparities, which will be revealed as simulations of the human vocal tract become more sophisticated.

### A. Future work

While the present study simulated the effects of gravity and myoelasticity on flap movement sequences, it omitted a third potentially important factor: Aerodynamics. Famously, phonation and trills are produced based on a combination of myoelastic principles combined with aerodynamic factors (Van Den Berg, 1958). However, the degree to which aerodynamic forces influence articulation during other speech acts must not be underestimated. Houde (1968), Perkell (1969), and Kent and Moll (1972), for example, noticed a forward looping of the tongue during the production of alveolar and velar stops; Hoole *et al.* (1998) later demonstrated

that aerodynamic forces influence the shape and extent of this forward looping. It is reasonable to assume that similar effects will take place during T production. Similarly, air-flow out of the mouth will produce forward and downward pressure on the tongue tip in roughly the direction of the  $[r^{\sim}]$ , an observation that can be seen directly in studies of airflow leaving the mouth during speech (Derrick *et al.*, 2009). We hope to follow up on this in future work.

### ACKNOWLEDGMENTS

This research was funded by a Discovery Grant from the Natural Sciences and Engineering Council of Canada (NSERC) to the second author, and by National Institutes of Health (NIH) Grant DC-02717 to Haskins Laboratories. Special thanks to Aislin Stott for labeling and segmenting the acoustic data.

- Barlow, S. M., and Estep, M. (2006). "Central pattern generation and the motor infrastructure for suck, respiration, and speech," *J. Commun. Disord.* **39**, 366–380.
- Boersma, P. (2001). "Praat, a system for doing phonetics by computer," *Glot Int.* **5**(9/10), 341–345. Available from <http://www.praat.org/> (Last viewed 5 November 2013).

- Browman, C. P., and Goldstein, L. (1986). "Towards an articulatory phonology," *Phonol. Yearbook* 3, 219–252.
- Browman, C. P., and Goldstein, L. (1989). "Articulatory gestures as phonological units," *Phonology* 6, 201–251.
- Browman, C. P., and Goldstein, L. (1992). "Articulatory phonology: An overview," *Phonetica* 49, 155–180.
- Buchaillard, S., Perrier, P., and Payan, Y. (2009). "A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning," *J. Acoust. Soc. Am.* 126(4), 2033–2051.
- Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row, New York), 470 pp.
- Cooper, F. S., and Abramson, A. S. (1960). *A Pilot X-ray Film of English Articulations With Stretched Sound* (Haskins Laboratories and Columbia-Presbyterian Medical Center, New York).
- Delattre, P., and Freeman, D. (1968). "A dialect study of American R's by x-ray motion picture," *Linguistics* 44, 29–68.
- Derrick, D., Anderson, P., Gick, B., and Green, S. (2009). "Characteristics of air puffs produced in English 'pa': Experiments and simulations," *J. Acoust. Soc. Am.* 125(4), 2272–2281.
- Derrick, D., and Gick, B. (2011). "Individual variation in English flaps and taps: A case of categorical phonetics," *Can. J. Ling.* 56(3), 307–319.
- Dominici, N., Ivanenko, Y. P., Cappellini, G., d'Avella, A., Mondì, V., Cicchese, M., Fabiano, A., Silei, T., Di Paolo, A., Giannini, C., Poppele, R. E., and Lacquaniti, F. (2011). "Locomotor primitives in newborn babies and their development," *Science* 334, 997–999.
- Fels, S., Vogt, F., Gick, B., Jaeger, C., and Wilson, I. (2003). "User-centered design for an open source 3D articulatory synthesizer," in *Proceedings of the 15th International Congress of Phonetic Science (ICPHS)*, pp. 179–182.
- Folkens, J. W., and Abbs, J. H. (1975). "Lips and jaw motor control during speech: Responses to resistive loading of the jaw," *J. Speech Hearing Res.* 18, 207–220.
- Gérard, J.-M., Perrier, P., and Payan, Y. (2006). "3D biomechanical tongue modeling to study speech production," in *Speech Production: Models, Phonetic Processes and Techniques*, edited by J. Harrington and N. Y. M. Tabain (Psychology Press, Sydney, Australia), pp. 85–102.
- Gick, B., and Stavness, I. (2013). "Modularizing speech," *Front. Psychol.* 4, 977.
- Grillner, S. (2006). "Biological pattern generation: The cellular and computational logic of networks in motion," *Neuron* 52, 751–766.
- Hannam, A., Stavness, I., Lloyd, J. E., and Fels, S. (2008). "A dynamic model of jaw and hyoid biomechanics during chewing," *J. Biomech.* 41(5), 1069–1076.
- Hoole, P., Munhall, K., and Mooshammer, C. (1998). "Do airstream mechanisms influence tongue movement paths?," *Phonetica* 55(3), 131–146.
- Houde, R. A. (1968). "A study of tongue body motion during selected speech sounds," in *Monogr. 2* (Speech Communication Research Laboratory, Santa Barbara, CA), 161 pp.
- Irtel, H. (2007). "PXLab: The Psychological Experiments Laboratory," Version 2.1.11, University of Mannheim, Mannheim, Germany, available from <http://www.pxlab.de> (Last viewed 5 November 2013).
- Kent, R., and Moll, K. (1972). "Cinefluorographic analysis of selected lingual consonants," *J. Speech Hearing Res.* 15, 453–473.
- Lisker, L., and Abramson, A. S. (1964). "A cross-language study of voicing in initial stops: Acoustical measurements," *Word* 20, 384–422.
- Lloyd, J., Stavness, I., and Fels, S. (2012). "ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," in *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, edited by Y. Payan and A. Gefen (Springer-Verlag, Berlin), pp. 355–394.
- Lund, J. P., and Kolta, A. (2006). "Brainstem circuits that control mastication: Do they have anything to say during speech?," *J. Commun. Disord.* 39, 381–390.
- Medicines and Healthcare Products Regulatory Agency (2004). "Evaluation report," MHRA Tech. Rep. 03107 (MHRA, UK), 67 pp.
- Meyer, D. E., and Gordon, P. C. (1985). "Speech production: Motor programming of phonetic features," *J. Mem. Language* 24, 3–26.
- Miyawaki, K., Hirose, H., Ushijima, T., and Sawashima, M. (1975). "A preliminary report on the electromyographic study of the activity of lingual muscles," *Ann. Bull. Res. Inst. Logopedics Phoniatrics* 9, 91–106.
- Mooney M. (1940). "A theory of large elastic deformation," *J. Appl. Phys.* 11(9), 582–592.
- Perkell, J. S. (1969). "Physiology of speech production: Results and implications of a quantitative cineradiographic study," in *Research Monogr. 53* (M.I.T. Press, Cambridge, MA), 120 pp.
- Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricaricof, R., and Zandipour, M. (2000). "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *J. Phonetics* 28, 233–272.
- Perrier, P., Payan, Y., Zandipour, M., and Perkell, J. (2003). "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study," *J. Acoust. Soc. Am.* 114(3), 1582–1599.
- R Core Team (2013). "R: A language and environment for statistical computing," R Foundation for statistical computing, Vienna, Austria, available at <http://www.R-project.org/> (Last viewed 5 November 2013).
- Rivlin, R. (1948). "Large elastic deformations of isotropic materials. iv. further developments of the general theory," *Philos. Trans. R. Soc. London Ser. A* 241(835), 379–397.
- Saltzman, E., and Byrd, D. (2000). "Task-dynamics of gestural timing: Phase windows and multifrequency rhythms," *Human Movement Sci.* 19, 499–526.
- Saltzman, E., and Kelso, J. A. S. (1987). "Skilled actions: A task-dynamic approach," *Psychol. Rev.* 94(1), 84–106.
- Shiller, D. M., Ostry, D. J., and Gribble, P. L. (1999). "Effects of gravitational load on jaw movements in speech," *J. Neurosci.* 19(20), 9073–9080.
- Slaughter, K., Li, H., and Sokoloff, A. J. (2005). "Neuromuscular organization of the superior longitudinalis muscle in the human tongue. I. Motor endplate morphology and muscle fiber architecture," *Cells Tissues Organs* 181, 51–64.
- Sloetjes, H., and Wittenburg, P. (2008). "Annotation by category—ELAN and ISO DCR," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, available from <http://tla.mpi.nl/tools/tla-tools/elan/> (Last viewed 5 November 2013).
- Stavness, I., Gick, B., Derrick, D., and Fels, S. (2012). "Biomechanical modeling of English /r/ variants," *J. Acoust. Soc. Am.* 131(5), EL355–EL360.
- Takemoto, H. (2001). "Morphological analysis of the human tongue musculature for three-dimensional modeling," *J. Speech Language Hearing Res.* 44, 95–107.
- Van Den Berg, J. W. (1958). "Myoelastic-aerodynamic theory of voice production," *J. Speech Hearing Res.* 1, 227–244.
- Westbury, J. R., Hashi, M., and Lindstrom, M. J. (1999). "Differences among speakers in lingual articulation for American English /ɹ/,  
Speech Commun. 26, 203–226.
- Zajac, F. E. (1988). "Muscle and tendon: Properties, models, scaling, and application to biomechanics and motor control," *Crit. Rev. Biomed. Eng.* 17(4), 359–411.
- Zue, W. V., and LaFerriere, M. (1979). "Acoustic study of medial /t,d/ in American English," *J. Acoust. Soc. Am.* 66(4), 1039–1050.